Tamanho Amostral Ótimo em Testes de Hipóteses Precisas

Patrícia Viana da Silva
Departamento de Estatística - Universidade de São Paulo

Victor Fossaluza
Departamento de Estatística - Universidade de São Paulo

Carlos Alberto de Bragança Pereira
Departamento de Estatística - Universidade de São Paulo

Resumo

Definição de tamanho de amostra é um assunto de grande interesse para a teoria estatística até mesmo como forma de avaliar a viabilidade de execução de um experimento. Essa avaliação é chamada de análise pré posteriori pela teoria Bayesiana, pois costuma ser realizada antes da obtenção dos dados e, nesse caso, tem como objetivo minimizar o custo de amostragem bem como a perda pela tomada dessa decisão. O objetivo desse trabalho é propor um método para determinar o n ótimo, tamanho de amostra, minimizando uma função de perda relacionada ao problema de testar hipóteses precisas.

1 Introdução

Em situações práticas nas quais se deseja realizar um experimento um dos primeiros questionamentos é o desenho e consequentemente o quanto se deve amostrar para obter a informação necessária à tomada de decisão. Sob a abordagem frequentista esse cálculo utiliza geralmente o intervalo e confiança. Já sob o paradigma bayesiano o tamanho amostral pode ser determinado via análise pré-posteriori tema tratado pela teoria da decisão.

A análise pré-posteriori, como o nome antecipa, é realizada antes da obtenção dos dados e, consequentemente, antes da obtenção da distribuição a posteriori e seu objetivo é escolher um experimento que minimize o custo geral dessa escolha ([1]). O custo geral é dado por uma função de perda que, nesse caso, consite da perda pela tomada da decisão e do custo de amostragem.

Nesse trabalho, estudaremos experimentos com a intensão de testar hipóteses precisas utilizando o Full Bayesian Statistical Test (FBST) proposto por [3]. Na próxima seção apresentare-

mos o procedimento do teste, bem como, uma função de perda adequada. Em seguida, na seção 3, os cálculos para a determinação do n ótimo e por fim, na seção 4, utilizamos um exemplo para ilustrar o método.

2 Testes de Hipóteses Precisas sob abodagem Bayesiana

Considere o modelo estatístico paramétrico, isto é, a quíntupla $(\mathcal{X}, \mathcal{A}, \mathcal{F}.\boldsymbol{\Theta}, \pi)$, em que $\mathcal{X}$ é o espaço amostral, $\mathcal{A}$ é a sigma-álgebra de subconjuntos do espaço paramétrico $\boldsymbol{\Theta}$ e π é a densidade a priori (sobre a sigma-álgebra de) $\boldsymbol{\Theta}$. Suponha um subconjunto $\boldsymbol{\Theta}_0$ de $\boldsymbol{\Theta}$ tendo medida de Lebesgue nula ($\boldsymbol{\Theta}_0 << \boldsymbol{\Theta}$) e seu complementar $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$. Seja, ainda, $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ a densidade a posteriori de $\boldsymbol{\theta}$ dada a observação amostral $\boldsymbol{x}$, $T(\boldsymbol{x}) = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|\boldsymbol{x}) > \sup_{\boldsymbol{\Theta}_0} \pi(\boldsymbol{\theta}|\boldsymbol{x})\}$ o conjunto tangente à $\boldsymbol{\Theta}_0$.

Para o espaço de decisões $D = \{Aceitar\ H_0(d_0), Rejeitar\ H_0(d_1)\}$ no caso em que se deseja testar $H_0 : \theta \in \boldsymbol{\Theta_0}(\theta = \theta_0)$, hipótese precisa, versus $H_1 : \theta \in \boldsymbol{\Theta}_1(\theta \neq \theta_0)$ podemos definir a seguinte função de perda:

Definição 1 *A função de perda L' em $D \times \boldsymbol{\Theta}$ em que $L'(Rejeitar\ H_0, \theta) = a[1 - \boldsymbol{I}_{(\theta \in T(\boldsymbol{x}))}]$ e $L'(Aceitar\ H_0, \theta) = b + c\boldsymbol{I}_{(\theta \in T(\boldsymbol{x}))}$.*

L' é uma função de perda que penaliza severamente a decisão por $H_0(d_0)$ quando θ é mais provável que o conjunto tangente a θ_0, $T(\boldsymbol{x})$. Madruga *et al* [2] demonstraram o teorema determinando que a função de perda L' é minimizada pelo procedimento de Pereira-Stern (FBST) apresentado a seguir:

Definição 2 *A medida de evidência de Pereira-Stern é definida como $EV(\boldsymbol{\Theta}_0, \boldsymbol{x}) = 1 - P(\boldsymbol{\theta} \in T(\boldsymbol{x})|\boldsymbol{x})$ e o procedimento do FBST consiste em aceitar $\boldsymbol{\Theta}_0$ quando $EV(\boldsymbol{\Theta}_0, \boldsymbol{x})$ for "grande" ([4]). No caso da função L', quando:*

$$EV(\boldsymbol{\Theta}_0, \boldsymbol{x}) > \frac{b+c}{a+c}$$

Assim, essa medida de evidência considera em favor da hipótese precisa todos os pontos do espaço paramétrico em que os valores da densidade a posteriori são, no máximo, tão grandes quanto o supremo sobre $\boldsymbol{\Theta}_0$, isso considera os pontos menos "prováveis" que algum ponto em $\boldsymbol{\Theta}_0$.

3 N ótimo

Seja a subsequência $\boldsymbol{X}^n = (X_1, X_2, \ldots, X_n)$ de variáveis aleatórias da sequência $\{X_i\}_{i \geq 1}$, que dado θ, são condicionalmente independentes e identicamente distribuidas com distribuição P_θ, bem como, a priori, $P(\theta)$ sendo a ditribuição de θ. Considere também o vertor $\boldsymbol{x}^n =$

$(x_1, x_2, \ldots, x_n)$, uma realização de $\boldsymbol{X}$. Nosso problema consite em encontrar o tamanho amostral ótimo n^* que minimiza o custo/perda geral esperada, dada por:

$$L(\theta, d, n) = L'(\theta, d) + C(n)$$

em que $L'(\theta, d)$ é a função de perda pela tomada da decisão, no caso, definida na seção anterior sendo minimizada pelo procedimento FBST e $C(n)$ é uma função de n, monótona, representando o custo ao observar $\boldsymbol{x}^n$.

E a perda esperada:

$$E[L(d, \theta, n)] = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\mathcal{X}}} [L'(\theta, d)] dP_\theta dP(\theta) + C(n)$$

Um caso particular, que utilizaremos no exemplo da próxima seção, considera um custo de amostragem constante $C(n) = kn$, em que $X|\theta \sim Binomial(n, \theta)$ com $P(X = x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, $\theta \sim Beta(1, 1)$ com $f(\theta) = \boldsymbol{I}_{\{\theta \in [0,1]\}}$ $X \sim Uniforme\{0, 1, \ldots, n\}$, nesse caso, $P(X = x|\theta) = \frac{1}{n+1}$ e, consequentemente, $\theta|\boldsymbol{X} \sim Beta(x+1, n-x+1)$ com $f(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \theta^x (1-\theta)^{n-x}$.

Dessa forma, a perda esperada é descrita abaixo:

$$\begin{aligned}
E_n[L((\theta, x), (n, \delta))] &= \int_{\boldsymbol{\Theta}} \int_{\mathcal{X}} L(\theta, d, n) f(x|\theta) f(\theta) dx d\theta + C(n) \\
&= \int_{\mathcal{X}} \left[\int_{\boldsymbol{\Theta}} L(\theta, d, n) f(\theta|x) d\theta \right] f(x) dx + C(n) \\
&= \int_{\mathcal{X}} \Big[\int_{T(x)} (b + c) \boldsymbol{I}_{\{\delta(x) = d_0\}} f(\theta|x) d\theta \\
&\quad + \int_{T(x)^c} [a + (b - a) \boldsymbol{I}_{\{\delta(x) = d_0\}} f(\theta|x) d\theta] \Big] dF(x) + C(n) \\
&= \int_{\mathcal{X}} \big[(b + c) \boldsymbol{I}_{\{\delta(x) = d_0\}} (1 - EV(x)) + aEV(x) \\
&\quad + (b - a) \boldsymbol{I}_{\{\delta(x) = d_0\}} EV(x) \big] dF(x) + C(n) \\
&= \tfrac{1}{n+1} \textstyle\sum_{x=0}^{n} [(b + c) - (a + c) EV(x)] \boldsymbol{I}_{\{EV(x) > \frac{b+c}{a+c}\}}
\end{aligned}$$

Assim, o tamanho amostral é determinado pelo valor n^*, ou seja, que minimiza essa perda esperada:

$$E_n[L(\theta, d, n)] = \frac{1}{n+1} \sum_{x=0}^{n} [(b + c) - (a + c) EV(x)] \boldsymbol{I}_{\{EV(x) > \frac{b+c}{a+c}\}}$$

4 Exemplo

Como exemplo, estamos interessados em testar:

$$\begin{cases} H_0 : p = 1/2, \\ H_1 : p \neq 1/2. \end{cases}$$

Nas condições citadas no final do capítulo anterior. Para o mesmo problema utilizamos alguns custos de observação diferentes.

A figura 1 apresenta a função de perda geral com custo por observação constante $C(n) = (0,7)n$. Observamos que o custo quando não há observação, $n = 0$, acaba sendo muito menor como é razoável supor.

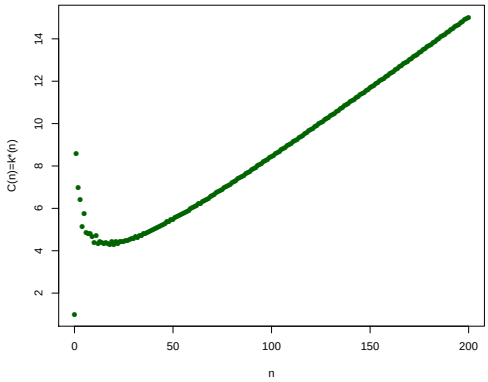

Figura 1: Custo constante.

Na figura 2, foi utilizado na função de perda um custo com base na função logarítmica natural $C(n) = ln(n)$ como custo por observação.

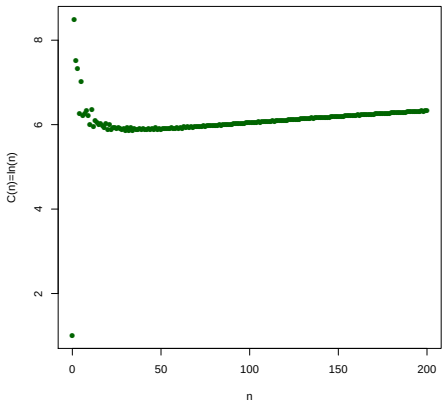

Figura 2: Custo logarítmico.

As rotinas de simulação e os gráficos foram desenvolvidas no programa R (http://www.r-project.org/).

5 Conclusões e Discussão

É possível construir, com a abordagem de teoria da decisão, um procedimento para determinação de tamanho amostral para este problema de teste de hipóteses precisas e também para muitos outros, no entanto, nem sempre de forma analítica.

Como extensão pretendemos estudar a aplicação do método para problemas de análise sequencial, em que, o processo de amostragem acaba quando uma decisão é tomada a respeito de um determinado problema.

Referências

[1] J. O. Berger, *Statistical decision theory and bayesian analysis*, 2 ed., Springs-Verlag, 1980.

[2] M. R. Madruga, L.G. Esteves, and S. Wechsler, *On the bayesianity of pereira-stern tests*, Sociedad de Estadística e investigacion Opcrativa **10** (2001), no. 2, 291–299.

[3] C. A. B. Pereira and J. M. Stern, *Evidence and credibility: Full bayesian significance test for precise hypotheses*, Entropy **1** (1999), no. 4, 99–110.

[4] C. A. B. Pereira, J. M. Stern, and S. Wechsler, *Can a significance test be genuinely bayesian?*, Bayesian Analysis **3** (2008), no. 1, 79–100.